Freie Universität Berlin

# Lecture: Machine Learning for Data Science

Winter semester 2021/22

## Lecture 1: Introduction

Prof. Dr. Eirini Ntoutsi

# Outline

- Why to study Machine Learning/Data Science?
- Why we need Machine Learning?
- What is Machine Learning/ Data Science?
- Main (machine) learning tasks
- Course content & logistics
- Things you should know from this lecture & reading material

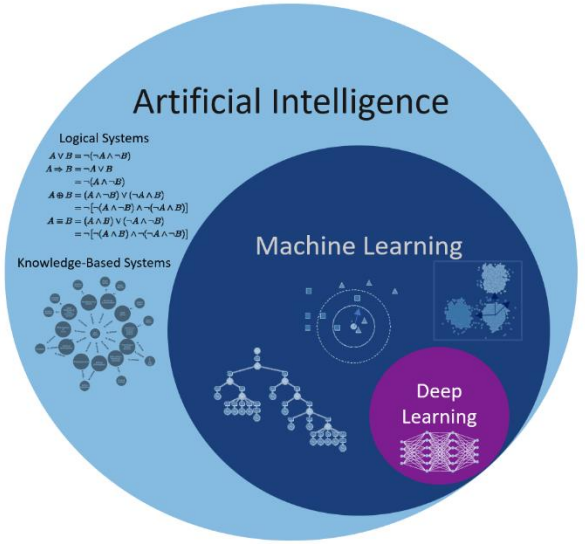
# Why to study Machine Learning – famous quotes

- "A breakthrough in machine learning would be worth ten Microsofts" (Bill Gates, Chairman, Microsoft) *
- "Machine learning is the next Internet" (Tony Tether, Director, DARPA) *
- "Machine learning is the hot new thing" (John Hennessy, President, Stanford) *
- "Web rankings today are mostly a matter of machine learning" (Prabhakar Raghavan, Dir. Research, Yahoo) *
- "Machine learning is going to result in a real revolution" (Greg Papadopoulos, Former CTO, Sun) *
- "Machine learning today is one of the hottest aspects of computer science" (Steve Ballmer, CEO, Microsoft) *
- "Just as the Industrial Revolution freed up a lot of humanity from physical drudgery, I think AI has the potential to free up humanity from a lot of the mental drudgery" (Andrew Ng, Coursera etc)

*Source: Pedro Domingos http://courses.cs.washington.edu/courses/cse446/15sp/slides/intro.pdf*

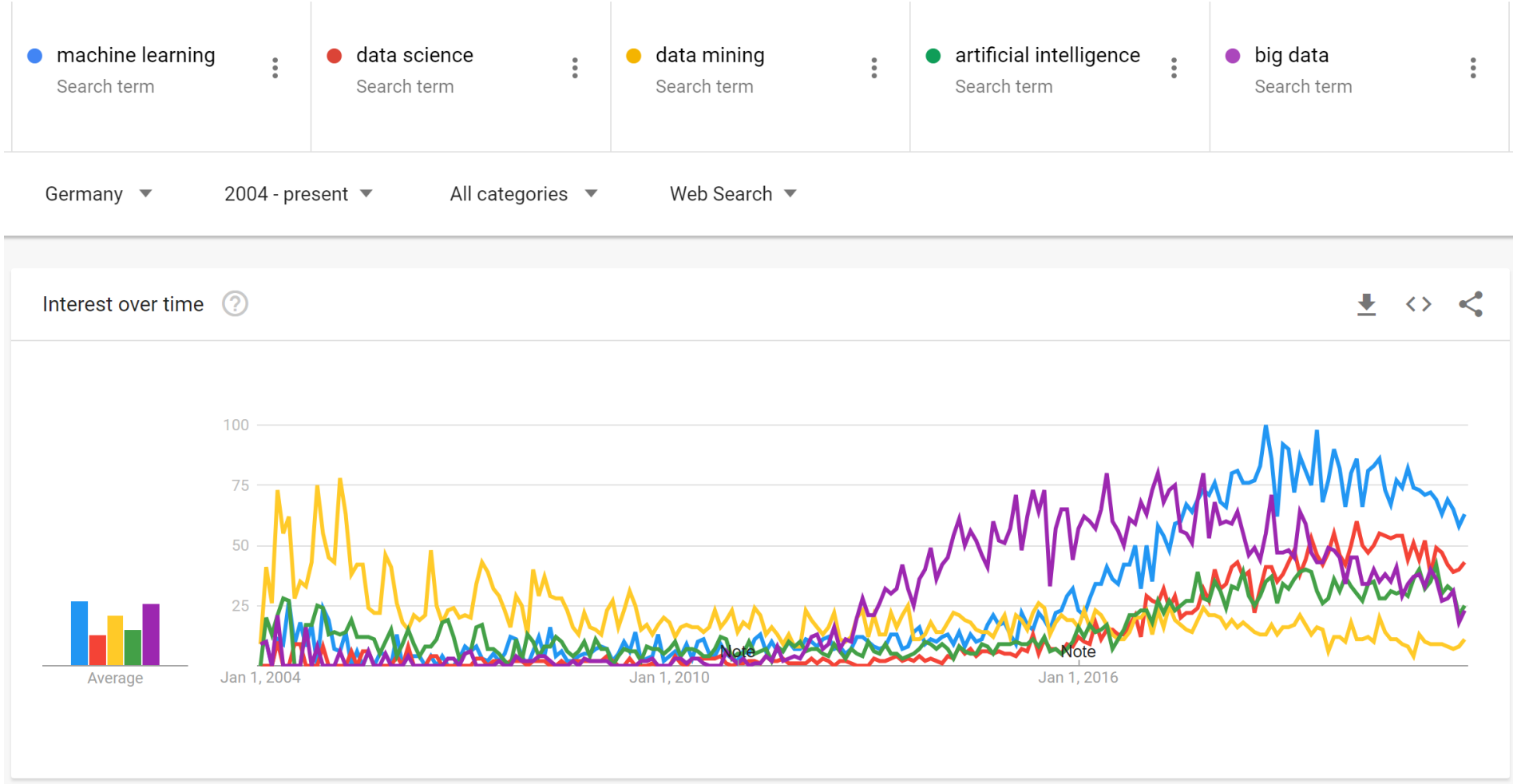
# Machine Learning, Deep Learning, Data Science, Data Mining, Big Data

- Machine Learning, data mining have been focusing on knowledge discovery from data for decades
  - Well defined set of tasks and solutions
- Big data and analytics are more business terms and ill-defined
  - *"Big data is like teenage sex: everyone talks about it, nobody really knows how to do it, everyone thinks everyone else is doing it, so everyone claims they are doing it."* Dan Ariely, Duke University
- "*Artificial Intelligence is intelligence demonstrated by machines, as opposed to the natural intelligence displayed by humans and animals*" Wikipedia
  - Machine Learning is a subset of AI
  - Deep learning is a subset of ML
- Disclaimer: In this lecture I will use the terms machine learning, data mining, data science (sometimes also Artificial Intelligence) interchangeably referring tc data-driven learning.




Source: https://data-science-blog.com/blog/2018/05/14/machine-learning-vs-deep-learning-wo-liegt-der-unterschied/

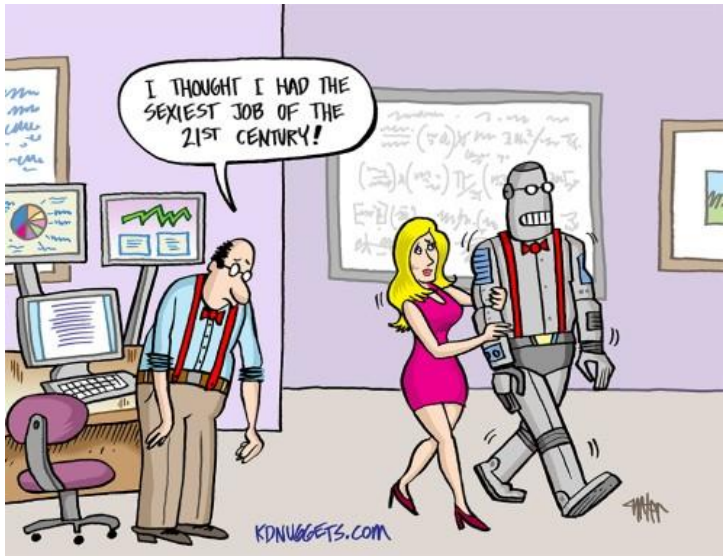
# Ever increasing interest & the ``rebranding'' effect

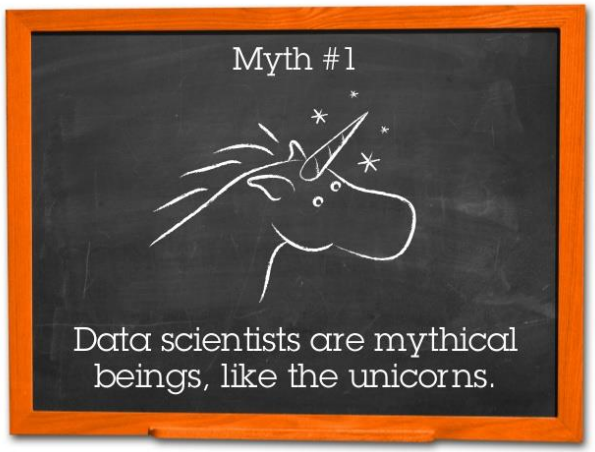



*Source: Google trends, query on 29.9.2021*

# Why to be a data scientist: The sexiest job of 21st century

*"If "sexy" means having rare qualities that are much in demand, data scientists are already there. They are difficult and expensive to hire and, given the very competitive market for their services, difficult to retain. There simply aren't a lot of people with their combination of scientific background and computational and analytical skills." Harvard Business Review. Data Scientist: The Sexiest Job of the 21st Century. October 2012 link*


Source: https://www.slideshare.net/IBMBDA/myths-and-mathemagical-superpowers-of-data-scientists

# The circumstances are great

- Data enablers
  - Web
  - Internet of things
  - Data intensive science
  - Big data
- (Intelligent) technology enablers
  - Mature DM, ML, AI
  - Deep learning
- Infrastructure enablers
  - Hardware advances
  - Software advances
- Everyone wants to join

Data deluge


Machine Learning advances

Computer power

Participation

# The democratization of AI

- "AI democratization is the spread of AI development to a wider user base that includes those without specialized knowledge of AI. The trend is being driven by large companies that are heavily invested in artificial intelligence, including IBM, Amazon, Facebook, Microsoft and Google, in the interests of furthering its development and adoption.
- AI development has typically demanded a lot of resources including expert-level knowledge, computing power and money. AI democratization involves facilitating development by providing user-friendly resources and supports such as pre-built algorithms, intuitive interfaces and high-performance cloud computing platforms. Having those supports in place makes it feasible for in-house developers without special expertise to create their own machine learning applications and other AI software."
  - Source: *https://whatis.techtarget.com/definition/AI-democratization*
- Related initiatives: Data democratization
- Beware of the risks→ Responsible AI/ Responsible Data Science ….

*"*Democratization of AI, a double-edged sword", https://towardsdatascience.com/democratization-of-ai-de155f0616b5*

# World-wide competition on Artificial Intelligence (AI)

- From FORSCHUNGSGIPFEL 2019 - Künstliche Intelligenz – Innovationstreiber einer neuen Generation, 19/3/2019, Berlin
- Cédric Villani's talk on the geopolitics of AI
  There are 3 fierce competitions
  - Competition for human talent
  - Competition for infrastructure
  - Competition for data


- More info on
  - Track #fogipf19 in Twitter
  - Check videos online: http://www.forschungsgipfel.de/2019/videos

# Outline

- Why to study Machine Learning/Data Science?
- Why we need Machine Learning?
- What is Machine Learning/ Data Science?
- Main (machine) learning tasks
- Course content & logistics
- Things you should know from this lecture & reading material

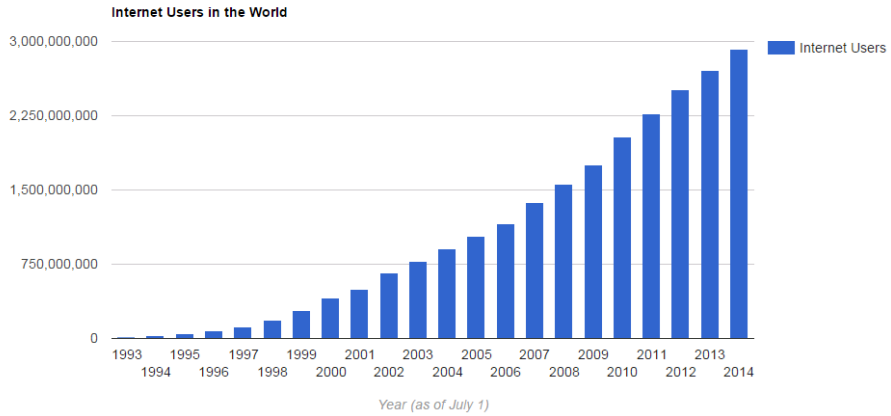
# Data everywhere

- Huge amounts of data are collected nowadays from different application domains
  - Google: processes 24 peta bytes of data per day.
  - Facebook: 10 million photos uploaded every hour.
  - Youtube: 1 hour of video uploaded every second.
  - Twitter: 400 million tweets per day.
  - Astronomy: Satellite data is in hundreds of PB.
  - ….

- The amount and the complexity of the collected data does not allow for manual analysis.

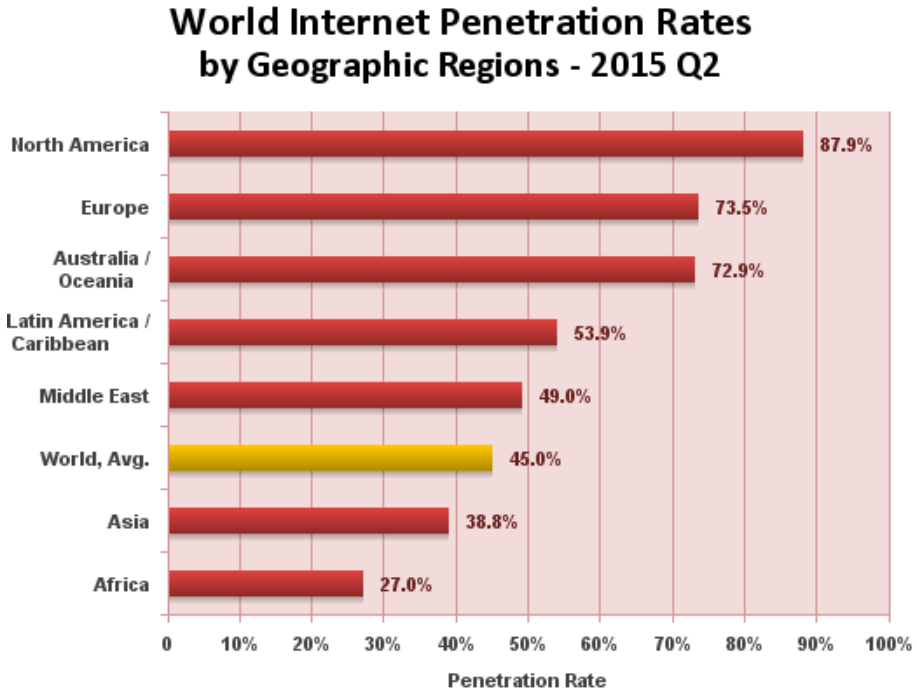
# Data sources: the Web

- Internet users




*Source: http://www.internetlivestats.com/internet-users/*


Web 2.0: A world of opinions
(user generated content)




Source: Internet World Stats - www.internetworldststs.com/stats.htm
Penetration Rates are based on a world population of 7,260,621,118 and 3,270,490,584 estimated Internet users on June 30, 2015.

# Data sources: Internet of things

- “The Internet of Things (IoT) is the network of physical objects or "things" embedded with electronics, software, sensors, and network connectivity, which enables these objects to collect and exchange data”, Wikipedia

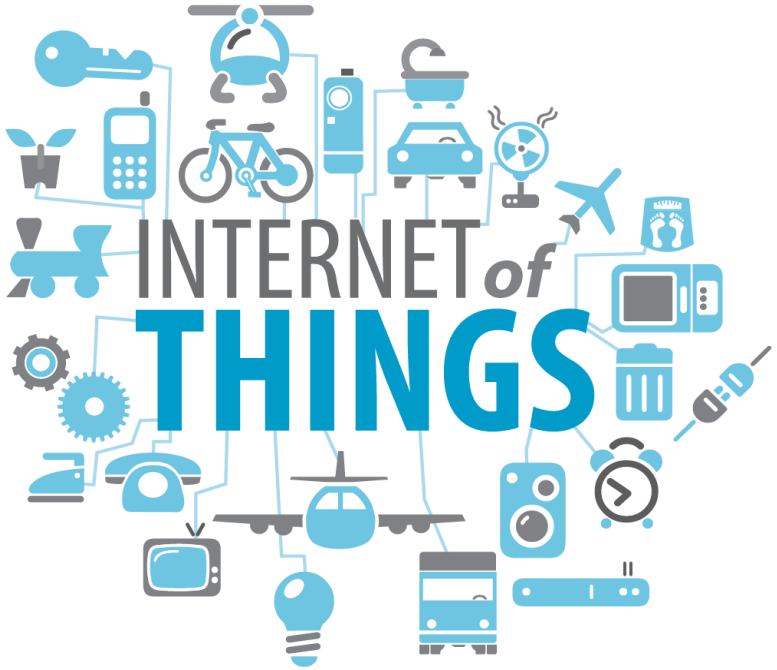



*Image source:http://tinyurl.com/prtfqxf*

During 2008, the number of things connected to the internet surpassed the number of people on earth… By 2020 there will be 50 billion … vs 7.3 billion people (2015).

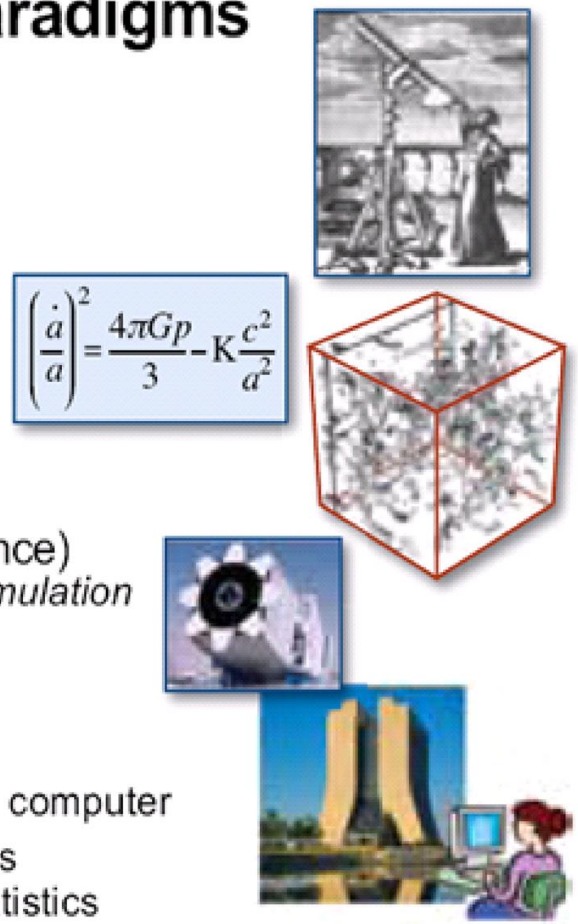
These things are everything, smartphones, tablets, refrigerators …. cattle.

*Source: http://blogs.cisco.com/diversity/the-internet-of-things-infographic*

# Data sources: Data intensive science

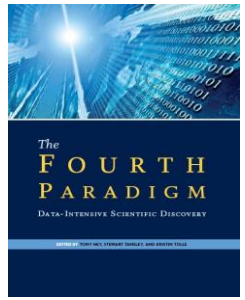
## Science Paradigms

- Thousand years ago:
  science was **empirical**
  *describing natural phenomena*
- Last few hundred years:
  **theoretical** branch
  *using models, generalizations*
- Last few decades:
  a **computational** branch
  *simulating complex phenomena*
- Today: **data exploration** (eScience)
  *unify theory, experiment, and simulation*
  - Data captured by instruments or generated by simulator
  - Processed by software
  - Information/knowledge stored in computer
  - Scientist analyzes database/files using data management and statistics




"Increasingly, scientific breakthroughs will be powered by advanced computing capabilities that help researchers manipulate and explore massive datasets."

*-The Fourth Paradigm – Microsoft Research*
*A tribute to Jim Gray*




Examples of e-science applications:

- Earth and environment
- Health and wellbeing
  - E.g., The Human Genome Project (HGP)
- Citizen science
- Scholarly communication
- Basic science
  - E.g., CERN

*Slide from:http://research.microsoft.com/en-us/um/people/gray/talks/nrc-cstb_escience.ppt*

# Data sources: Manufacturing

- Andrew Ng Says Factories Are AI's Next Frontier
  - Source: https://www.technologyreview.com/s/609770/andrew-ng-says-factories-are-ais-next-frontier/

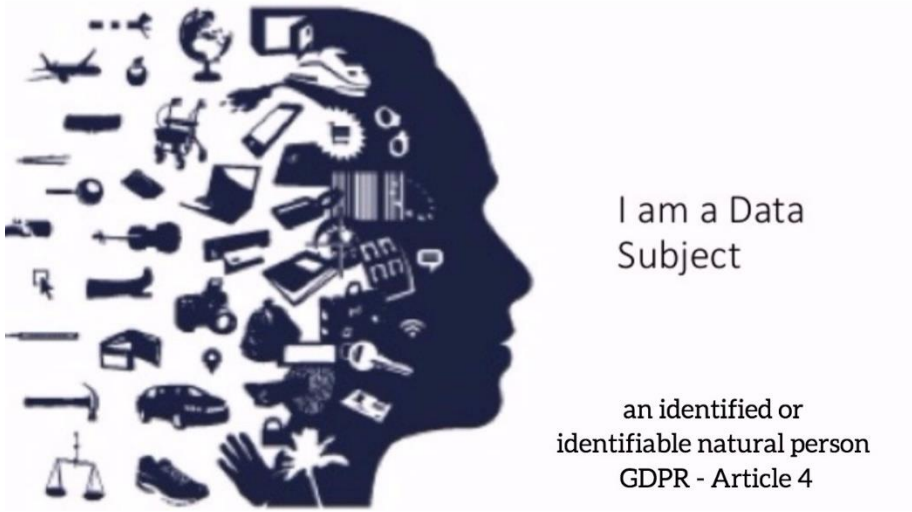
*Image source: https://images.readwrite.com/wp-content/uploads/2018/03/AAEAAQAAAAAAAAueAAAAJDY1NmFlN2NhLWExZTUtNDRhNy1iMWQ5LTViZGM3NTFlODczYQ.jpg*

Companies are making major investments in AI and industrial analytics to help drive their digital transformation

*Image source: https://cdn-sv1.deepsense.ai/wp-content/uploads/2018/04/Spot-the-flaw-Visual-quality-control-in-manufacturing-1140x337.jpg*

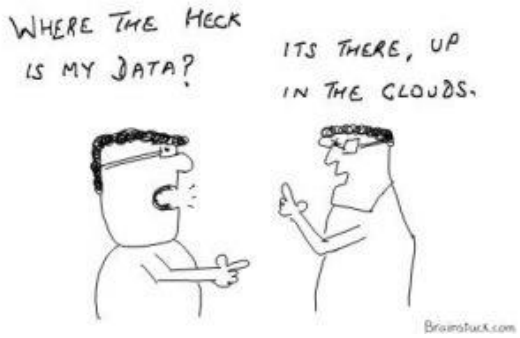
# Data sources: We … the data subjects

- Wherever we go, we are "datafied".
- Smartphones are tracking our locations.
- We leave a data trail in our web browsing.
- Interaction in social networks.




Source: https://abhisrivastava.com/2017/09/09/who-is-a-data-subject-in-gdpr/

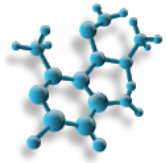
- Privacy is an important issue … not covered though in this lecture → privacy aware machine learning
  - EU General Data Protection Regulation (https://eugdpr.org/)



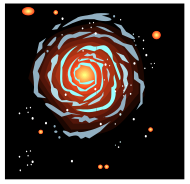

# From data to knowledge

- *"We are drowning in information but starving for knowledge"*
  John Naibett link

*telecommunications*



*healthcare*

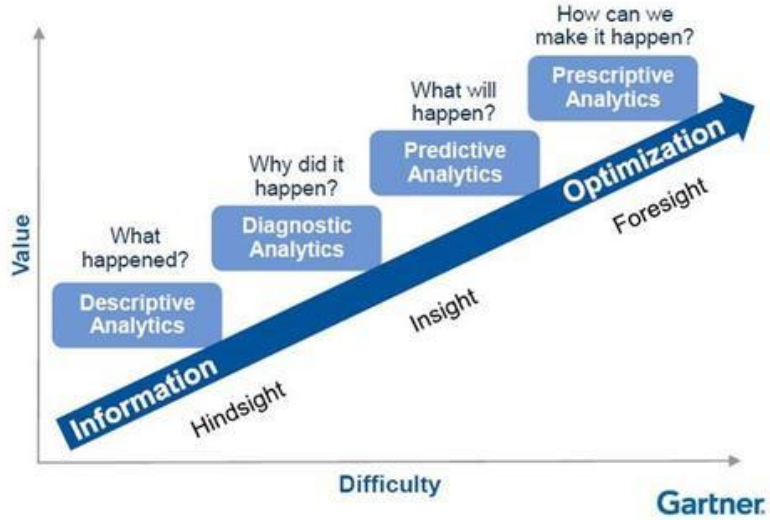


*astronomy*

*transactions*

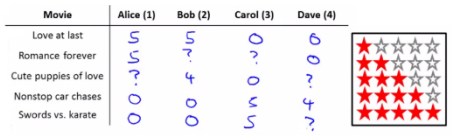

*Web*

- Manual analysis is infeasible → (semi)automatic ways to extract knowledge from the data
- Knowledge can take many forms
  - Descriptive modeling: Explains the characteristics and behavior of observed data.
  - Predictive modeling: Predicts the behavior of new data based on some model built upon historical data.
  - …




*Source: https://www.zdnet.com/article/data-to-analytics-to-ai-from-descriptive-to-predictive-analytics/*

# From data to knowledge of different types/forms

| Data sources | Methods | Knowledge |
|---|---|---|
|  Call records | Outlier Detection | Detect fraud cases |
|   Movie ratings | Collaborative filtering | Recommend movies to users |
|  Telescope images | Classification | Is it an «early», «intermediate» or «late formation» star? |
|  News articles | Clustering | What are the topics people discuss about in the news today? |

# Outline

- Why to study Machine Learning/Data Science?
- Why we need Machine Learning?
- What is Machine Learning/ Data Science?
- Main (machine) learning tasks
- Course content & logistics
- Things you should know from this lecture & reading material

# How do machines learn?

- ML "*gives computers the ability to learn without being explicitly programmed*" (Arthur Samuel, 1959)
- We don't codify the solution. We don't even know it!
- An example of explicitly programming a solution

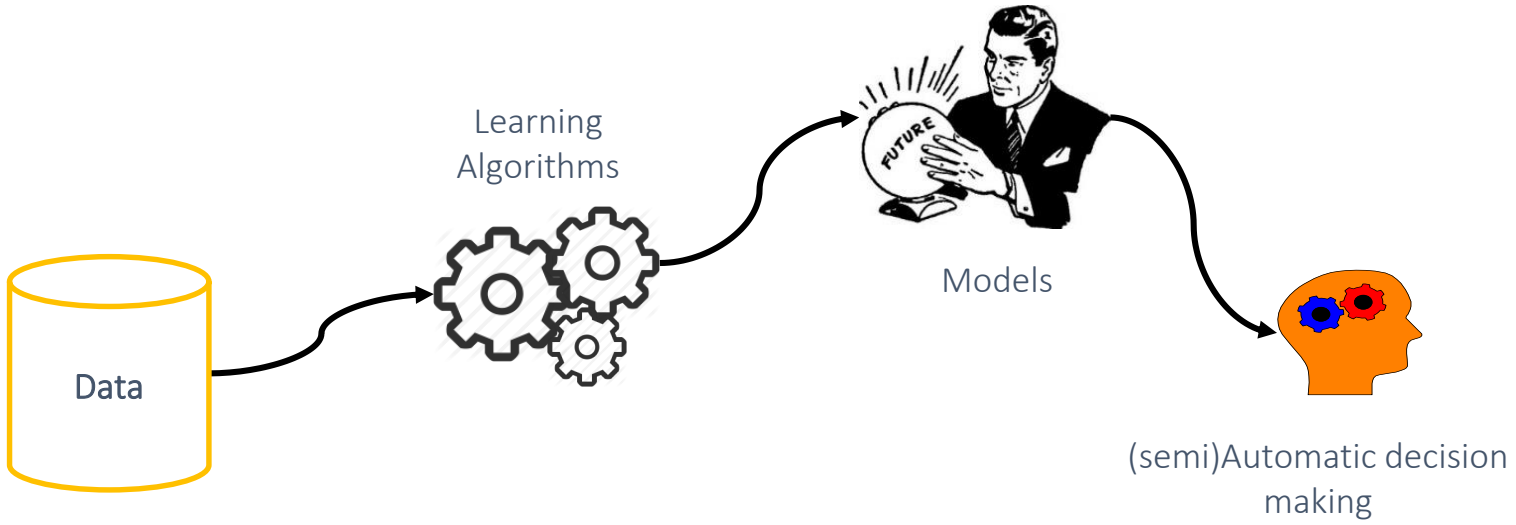
```
INSERTION-SORT(A)
1  for j ← 2 to length[A]
2      do key ← A[j]
3         ▷ Insert A[j] into the sorted sequence A[1 . . j − 1].
4         i ← j − 1
5         while i > 0 and A[i] > key
6             do A[i + 1] ← A[i]
7                i ← i − 1
8         A[i + 1] ← key
```


- Can you provide such an explicit algorithm for cat recognition?

*Source: https://en.wikipedia.org/wiki/Cat#/media/File:Cat_poster_1.jpg*

# How do machines learn?

- ML "*gives computers the ability to learn without being explicitly programmed*" (Arthur Samuel, 1959)
- We don't codify the solution. We don't even know it!
- Data is the key & the learning algorithm

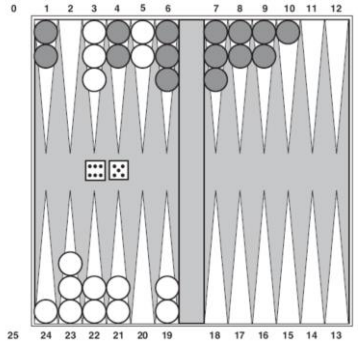


# More formally: How do machines learn?

- *"A computer program is said to learn from experience E w.r.t. some class of tasks T and performance measure P, if its performance at tasks in T, as measured by P, improves with experience E."*

  Tom Mitchell, Machine Learning 1997.

- Example: A backgammon learning problem
  - Task *T*: playing backgammon
  - Performance measure *P*: % of games won against opponents
  - Training experience *E*: playing practice games against itself
- Example: Exam performance
  - Task *T*: predict whether a student will pass the final exam or not
  - Experience *E*: historical records of students that took the exam
  - Performance measure *P*: % of correctly identified students



# (Machine) Learning from experience 1/2

- Experience comes in terms of data from the specific problem/ application
- Datasets consists of instances (also known as examples or objects)
  - e.g., in a university database: students, professors, courses, grades,…
  - e.g., in a movie database: movies, actors, director,…
- Instances are described through features (also known as attributes or variables)
  - E.g. a course is described in terms of a title, description, lecturer, teaching frequency etc.
- Tabular data representation (rows are the instances and the columns are the features)
  - Assumes all instances have the same feature representation

| ID | Gender | Height(cm) | Weight (kg) | Hair Color | Blood Group | Glasses | Smoker | GGS 787 Grade |
|---|---|---|---|---|---|---|---|---|
| 67 | Female | 175 | 60 | brown | A | no | frequent | A+ |
| 68 | Female | 176 | 52 | blond | AB | yes | frequent | A |
| 69 | Female | 176 | 63 | black | A | yes | casual | A+ |
| 70 | Female | 179 | 65 | brown | 0 | yes | no | B |

- But data are not necessarily tabular data
  - In this course we assume mainly tabular data with a fixed feature representation

# (Machine) Learning from experience 2/2

- Except for the instance description, we might also have feedback on those instances from some "teacher"/"expert"
  - E.g., whether a student passed the exam
- The direct feedback is known as class attribute/label, i.e., each instance is associated with a label → labeled dataset
- But we might have no feedback at all → unlabeled dataset
- There might be also indirect feedback

*Lecture 2 is devoted on getting to know our data!*

| fruit | length | width | weight | label |
|---|---|---|---|---|
| fruit 1 | 165 | 38 | 172 | Banana |
| fruit 2 | 218 | 39 | 230 | Banana |
| fruit 3 | 76 | 80 | 145 | Orange |
| fruit 4 | 145 | 35 | 150 | Banana |
| fruit 5 | 90 | 88 | 160 | Orange |
| … | | | | |
| fruit n | … | … | … | … |

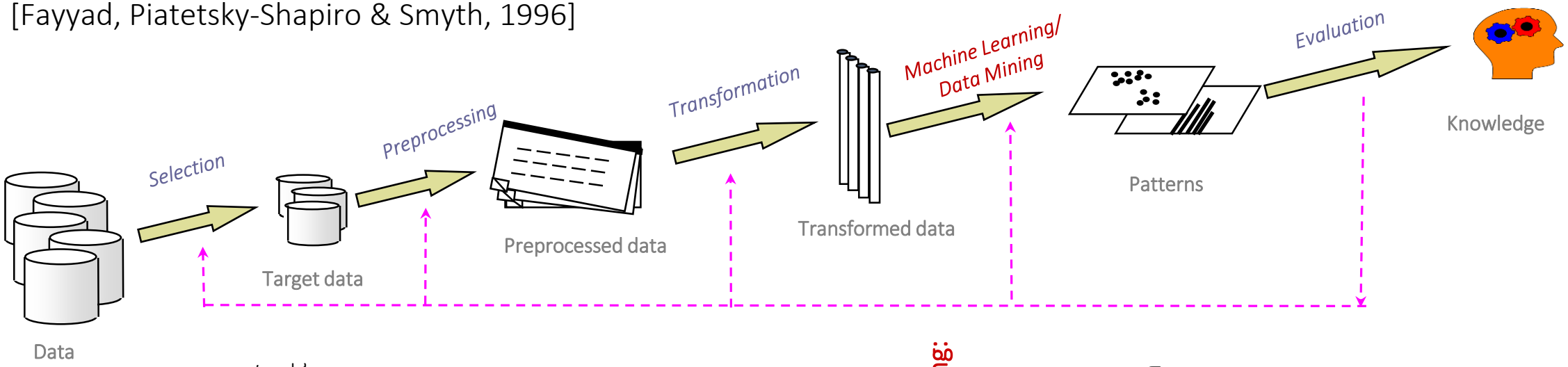
Labeled dataset

| fruit | length | width | weight |
|---|---|---|---|
| fruit 1 | 165 | 38 | 172 |
| fruit 2 | 218 | 39 | 230 |
| fruit 3 | 76 | 80 | 145 |
| fruit 4 | 145 | 35 | 150 |
| fruit 5 | 90 | 88 | 160 |
| … | | | |
| fruit n | … | … | … |

Unlabeled dataset

# Machine Learning is not the only step

[Fayyad, Piatetsky-Shapiro & Smyth, 1996]




**Selection:**
- Select a relevant dataset or focus on a subset of a dataset
- File / DB/

**Preprocessing/Cleaning:**
- Integration of data from different data sources
- Noise removal
- Missing values

**Transformation:**
- Select useful features
- Feature transformation/ discretization
- Dimensionality reduction

**Machine Learning/ Data Mining:**
- Search for patterns of interest

**Evaluation:**
- Evaluate patterns based on interestingness measures
- Statistical validation of the Models
- Visualization
- Descriptive Statistics

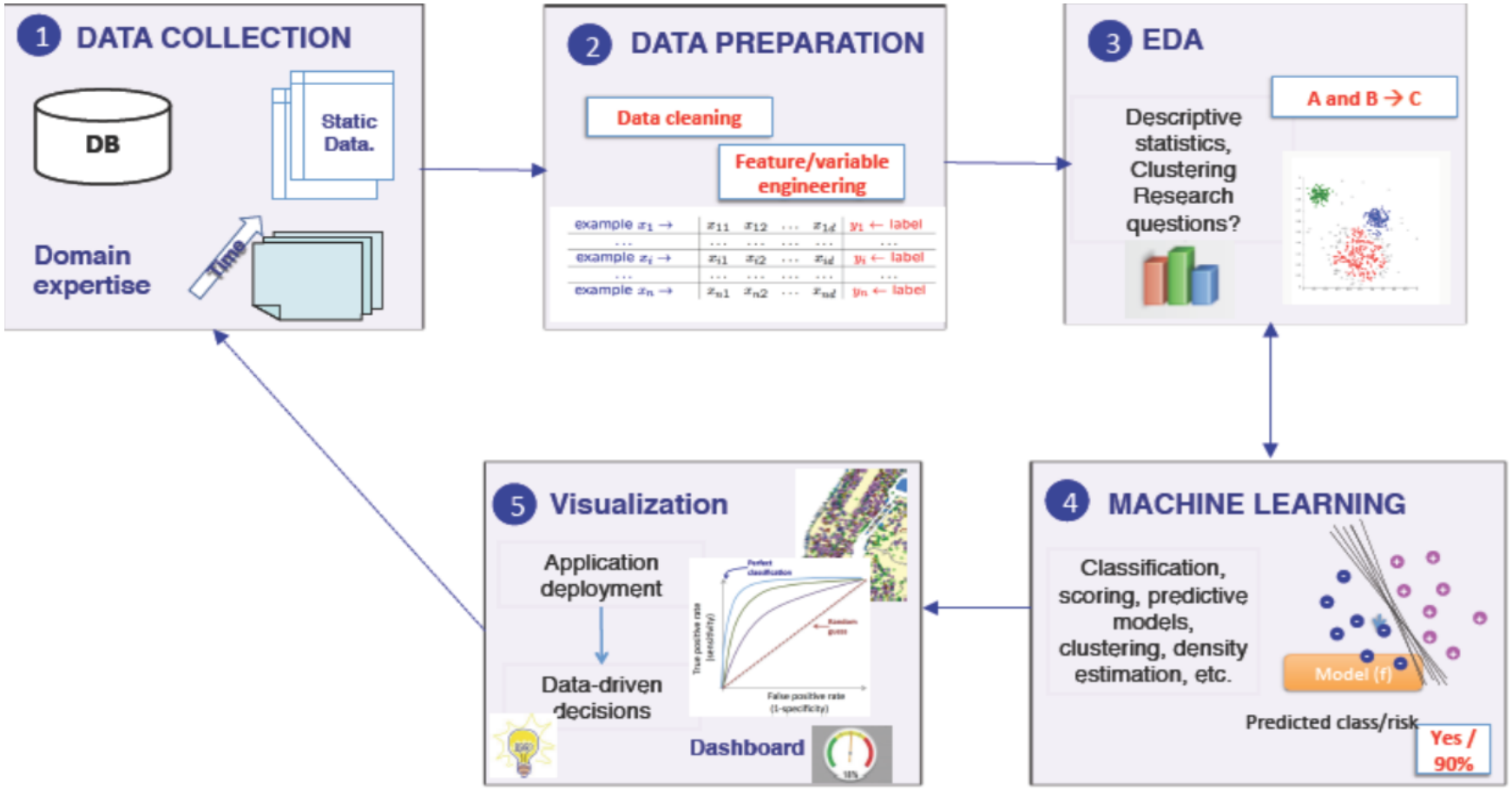
# The KDD process

- *"Knowledge Discovery in Databases (KDD) is the nontrivial process of identifying valid, novel, potentially useful, and ultimately understandable patterns in data."*, [Fayyad, Piatetsky-Shapiro, and Smyth 1996]

- Remarks:
  - valid: the discovered patterns should also hold for new, previously unseen problem instances.
  - novel: at least to the system and preferably to the user
  - potentially useful: they should lead to some benefit to the user or task
  - ultimately understandable: the end user should be able to interpret the patterns either immediately or after some post-processing

**Clarification:** The term databases does not refer exclusively to relational databases storing structured data … it can be any data storage and also structured, semi-structured, non-structured data

# A modern version: The Data Science process



# Outline

- Why to study Machine Learning/Data Science?
- Why we need Machine Learning?
- What is Machine Learning/ Data Science?
- Main (machine) learning tasks
- Course content & logistics
- Things you should know from this lecture & reading material

# Different learning tasks

| fruit | length | width | weight | label |
|---|---|---|---|---|
| fruit 1 | 165 | 38 | 172 | Banana |
| fruit 2 | 218 | 39 | 230 | Banana |
| fruit 3 | 76 | 80 | 145 | Orange |
| fruit 4 | 145 | 35 | 150 | Banana |
| fruit 5 | 90 | 88 | 160 | Orange |
| ... | | | | |
| fruit n | ... | ... | ... | ... |

- Based on the feedback we have on the data, we can distinguish between:
- Direct-feedback instances — Supervised learning
  - the correct response /label is provided for each instance by the "teacher"
  - e.g., good or bad product

| fruit | length | width | weight |
|---|---|---|---|
| fruit 1 | 165 | 38 | 172 |
| fruit 2 | 218 | 39 | 230 |
| fruit 3 | 76 | 80 | 145 |
| fruit 4 | 145 | 35 | 150 |
| fruit 5 | 90 | 88 | 160 |
| ... | | | |
| fruit n | ... | ... | ... |

- No-feedback instances — Unsupervised learning
  - no evaluation/label of the instance is provided, since there is no "teacher"
  - e.g., no information on whether a product is good or bad, just the description of the product/instance
- Indirect-feedback instances — Reinforcement learning
  - less feedback is given, since not the proper action, but only an evaluation of the chosen action is given by the teacher

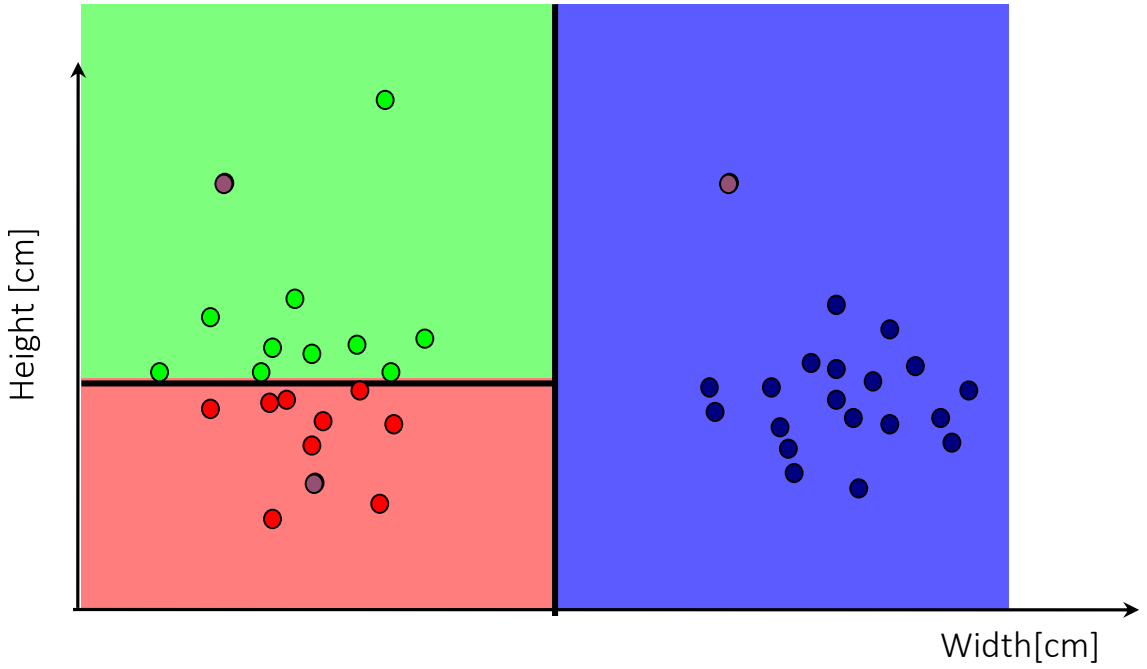
# Learning tasks: Supervised learning: Classification

- Supervised learning/ Predictive:
  - A description of the instances and their class labels is available
  - The class attribute is discrete
  - The goal is to learn a mapping from the instances to the class labels, i.e., given a future unseen instance to predict its class label

| fruit | length | width | weight | label |
|---|---|---|---|---|
| fruit 1 | 165 | 38 | 172 | Banana |
| fruit 2 | 218 | 39 | 230 | Banana |
| fruit 3 | 76 | 80 | 145 | Orange |
| fruit 4 | 145 | 35 | 150 | Banana |
| fruit 5 | 90 | 88 | 160 | Orange |
| ... | | | | |
| fruit n | ... | ... | ... | ... |

- Typical examples covered in this lecture:
  - Classification
  - Regression
  - Outlier detection

# Classification: an example




| instance | width | height | class |
|---|---|---|---|
| 1 | 2,6 | 4,5 | Screw |
| 2 | 3,7 | 7,3 | Nails |
| 3 | 4,1 | 6,5 | Paper Clips |
| 4 | 8,5 | 8,1 | Screw |
| 5 | 9,5 | 5,5 | Nails |
| … | … | … | … |

Screw

Nails

Paper clips

**New object**

- The goal is to learn a mapping from the "height, width space" to the class space (nails, screw,paper clips)
- For the new objects, the result of the classification if one of the class labels {nails, screw,paper clips}

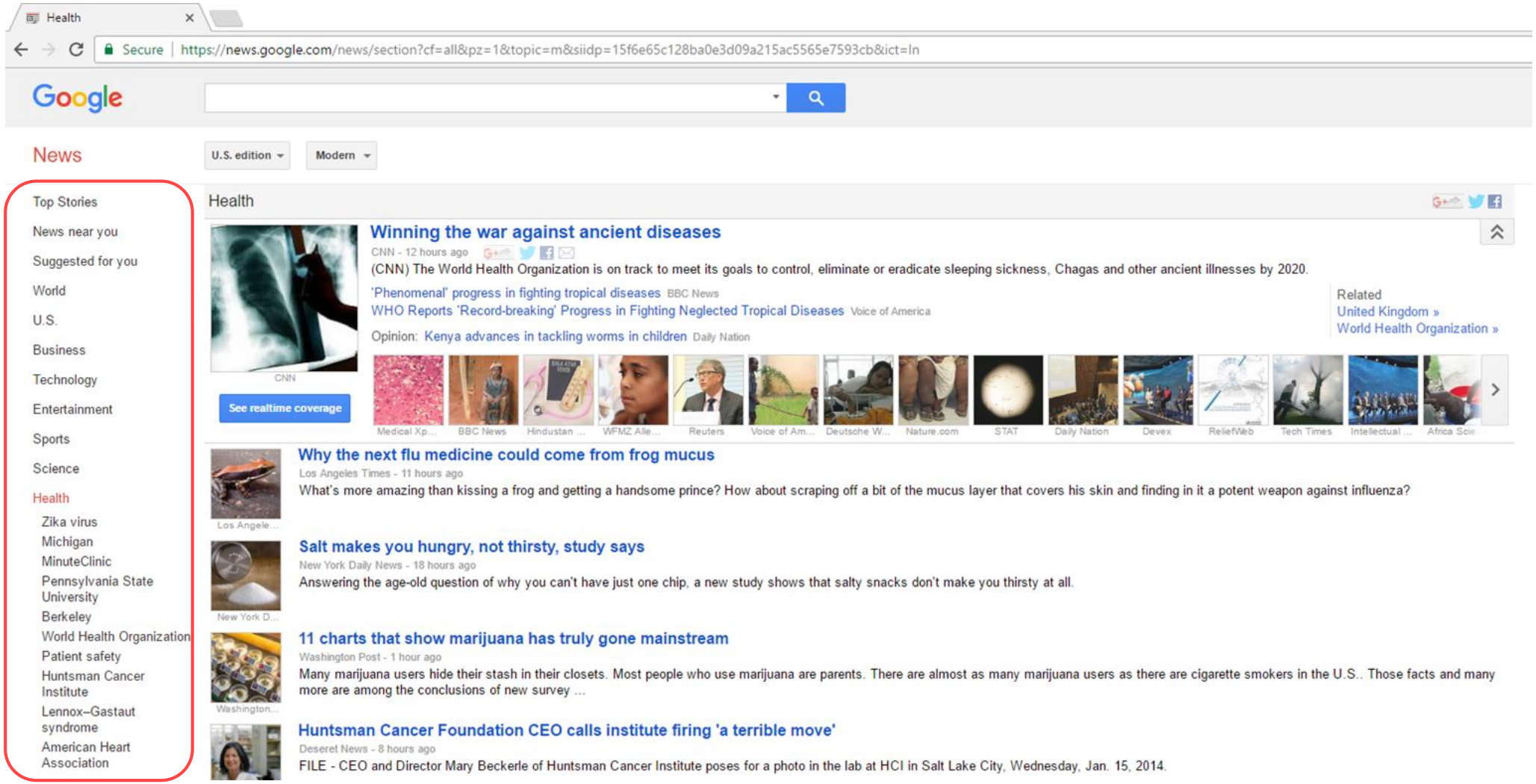
# Classification applications 1/3

- Application: Fraud Detection
  - Goal: Predict fraudulent cases in credit card transactions.
  - Approach:
    - Use credit card transactions and the information on its account-holder as attributes.
      - When does a customer buy, what does he buy, how often he pays on time, etc
    - Label past transactions as fraud or fair transactions. This forms the class attribute.
    - Learn a model for the class of the transactions.
    - Use this model to detect fraud by observing credit card transactions on an account.

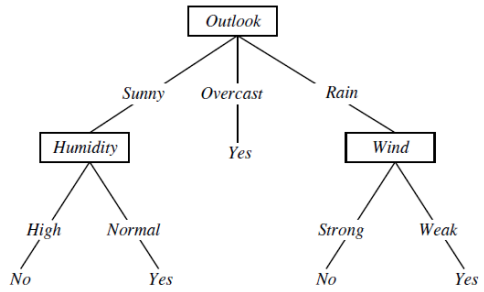
# Classification applications 2/3

- Application: Churn prediction in telco
  - Goal: Predict whether a customer is likely to be lost to a competitor
  - Approach:
    - Use detailed record of transactions with each of the past and present customers, to find attributes.
      - How often the customer calls, where he calls, what time-of-the day he calls most, his financial status, marital status, etc.
    - Label the customers as loyal or disloyal (class attribute).
    - Find a model for customer loyalty
    - Use this model to predict churn and organize possible retention strategies.

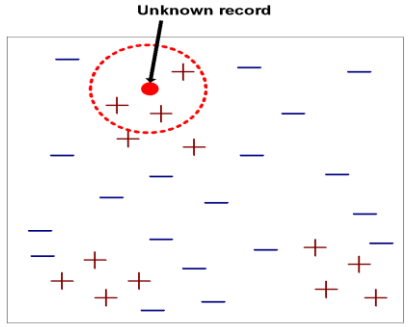
# Classification applications 3/3



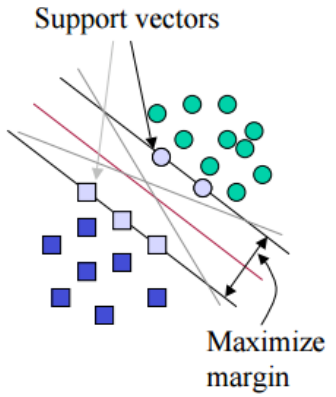
# A huge variety of classification algorithms



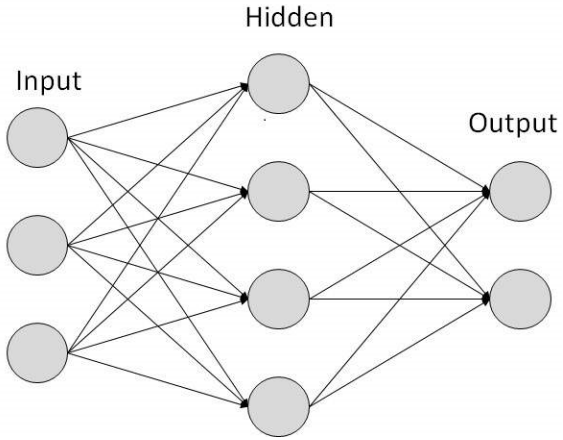

Decision trees



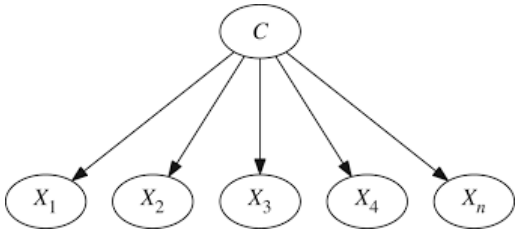

*k* nearest neighbours



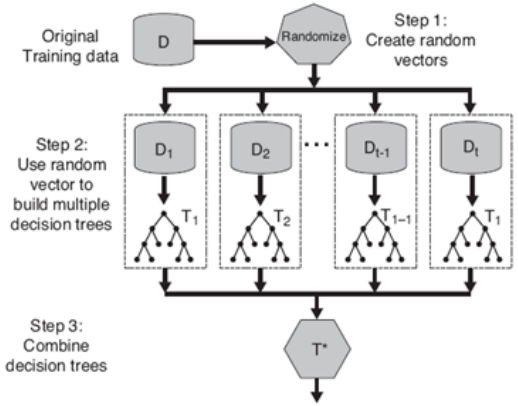

Support vector machines




Neural networks




Bayesian classifiers

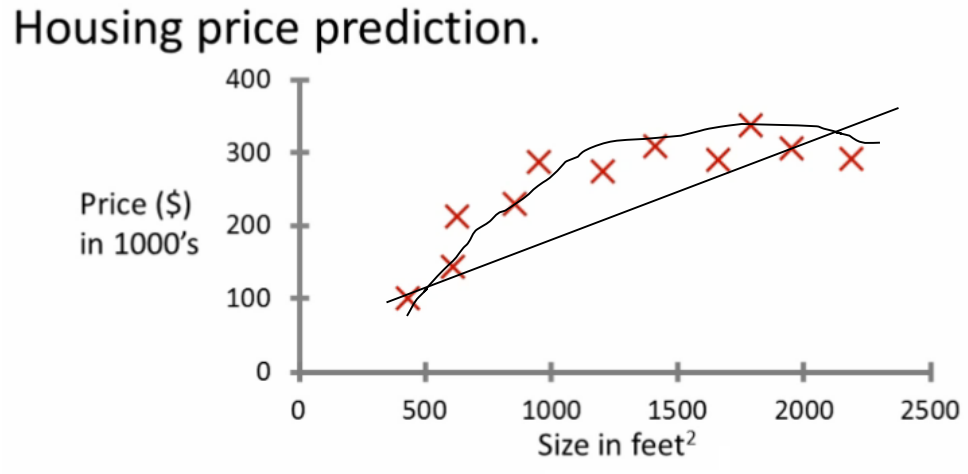



Ensembles

# Learning tasks: Supervised learning: Regression

- Similar to classification, but the class attribute is continuous rather than discrete.
- Goal: Predict a value of a given continuous valued variable based on the values of other variables.

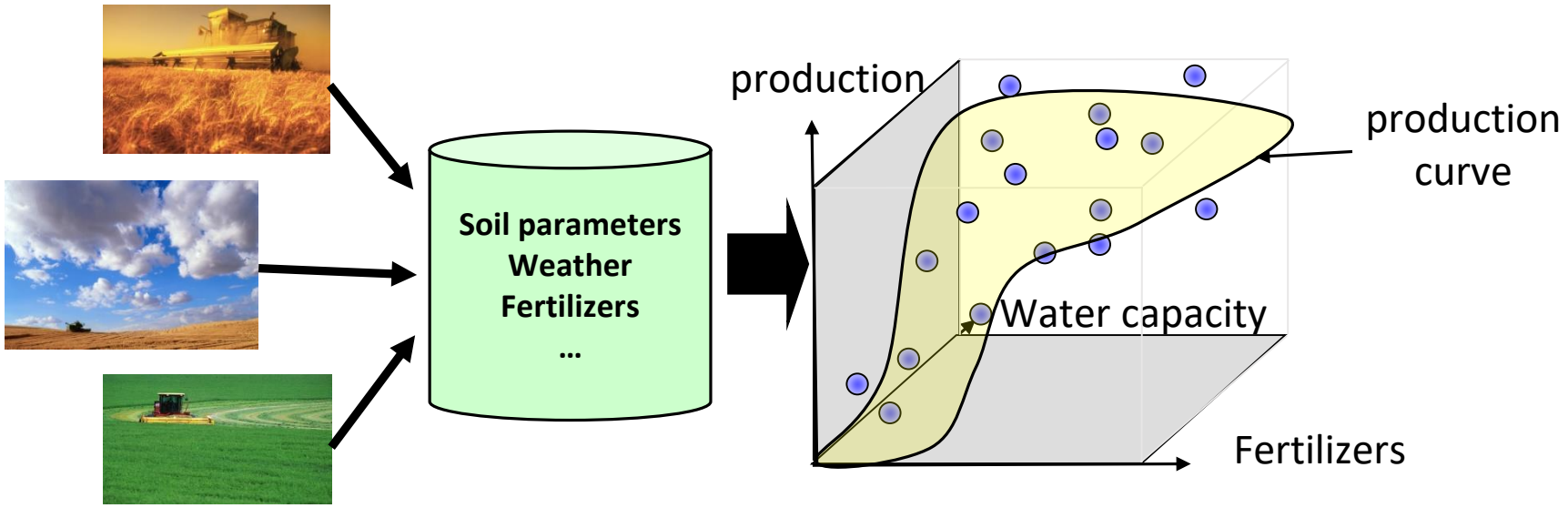
| instance | size | price |
|---|---|---|
| 1 | 500 | 100 |
| 2 | 1000 | 250 |
| 3 | 2000 | 300 |




Given this data, a friend has a house 750 square feet - how much can they be expected to get?

*Source: Andrew Ng ML course, Coursera*

# Regression application: Precision farming




- Create a production curve depending on multiple parameters like soil characteristics, weather, used fertilizers.
- Only the appropriate amount of fertilizers given the environmental settings (soil, weather) will result in maximum yield.
- Controlling the effects of over-fertilization on the environment is also important

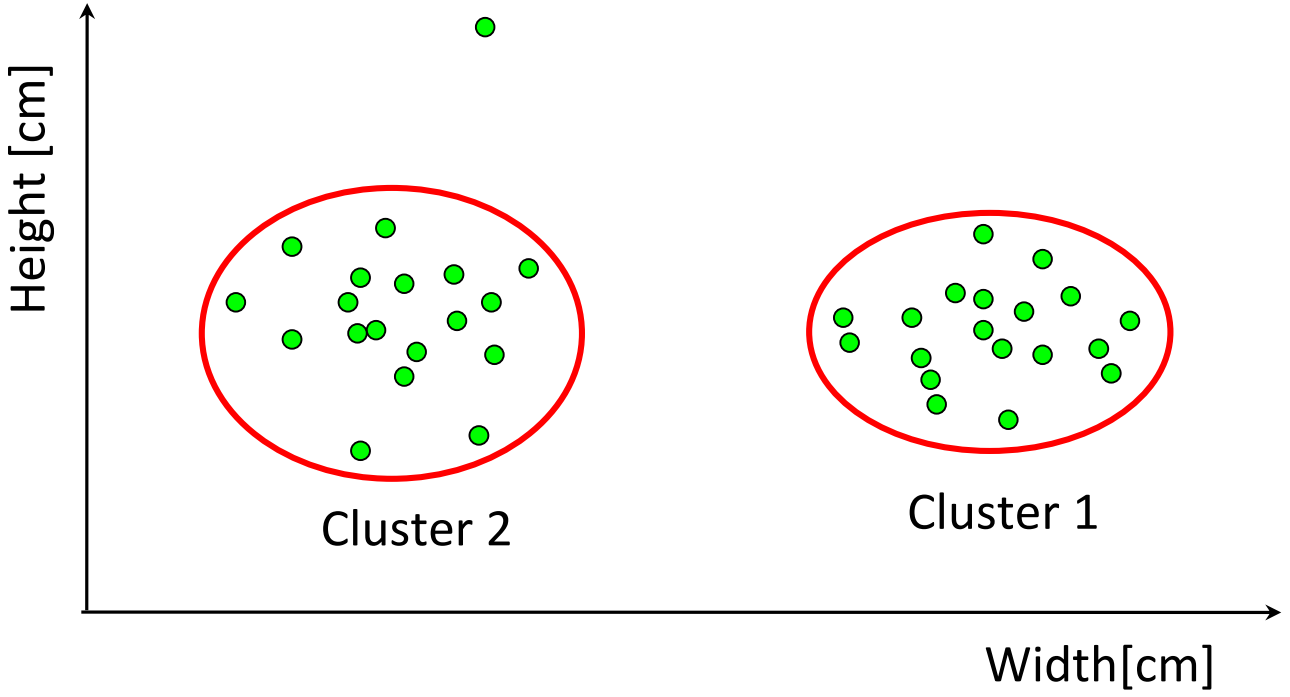
# Learning tasks: Unsupervised learning

- Unsupervised learning/ Descriptive:
  - Only a description of the instances is available
  - No feedback/labels/class attribute are available
  - The goal is to discover structure in the data

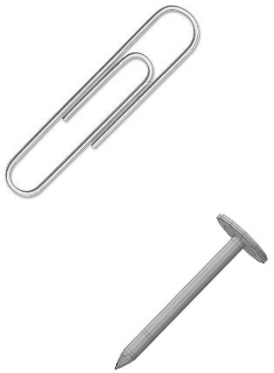
- Typical subtasks covered in this lecture:
  - clustering
  - outlier detection

| fruit | length | width | weight |
|---|---|---|---|
| fruit 1 | 165 | 38 | 172 |
| fruit 2 | 218 | 39 | 230 |
| fruit 3 | 76 | 80 | 145 |
| fruit 4 | 145 | 35 | 150 |
| fruit 5 | 90 | 88 | 160 |
| ... | | | |
| fruit n | ... | ... | ... |

# Clustering: an example




| instance | width | height |
|---|---|---|
| 1 | 2,6 | 4,5 |
| 2 | 3,7 | 7,3 |
| 3 | 4,1 | 6,5 |
| 4 | 8,5 | 8,1 |
| 5 | 9,5 | 5,5 |
| … | … | … |



- Each point described in terms of its height and width
- No information on the actual classes (nails, paper clips) is available to the clustering algorithm.

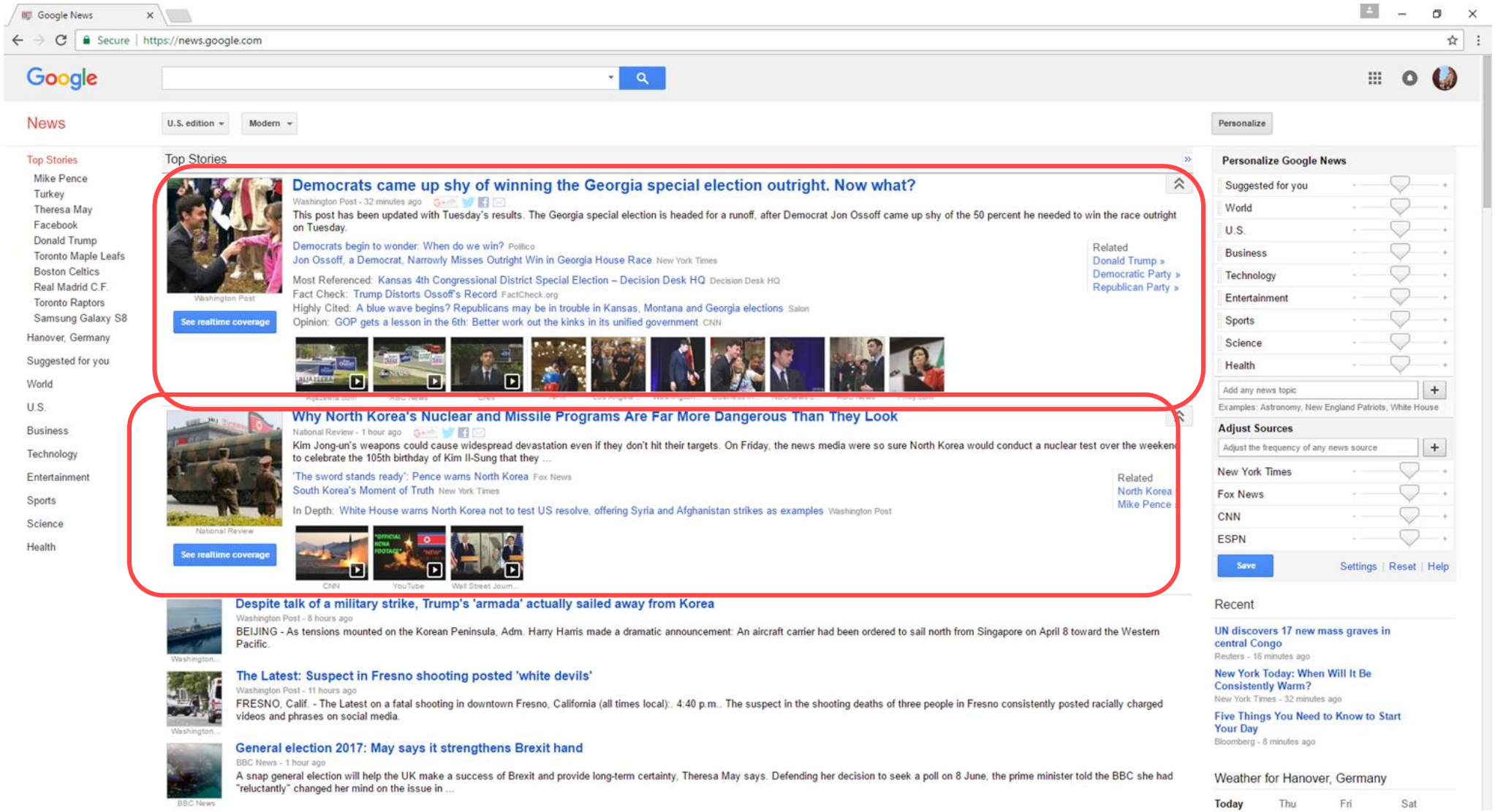
# Clustering applications 1/3

- Application: Market Segmentation
- Goal: subdivide a market into distinct subsets of customers where any subset may conceivably be selected as a market target to be reached with a distinct marketing mix.
- Approach:
  - Collect different attributes of customers based on their geographical and lifestyle related information.
    - E.g., age, income, education, family status, ….
  - Find clusters of similar customers.
  - Measure the clustering quality by observing buying patterns of customers in same cluster vs. those from different clusters.

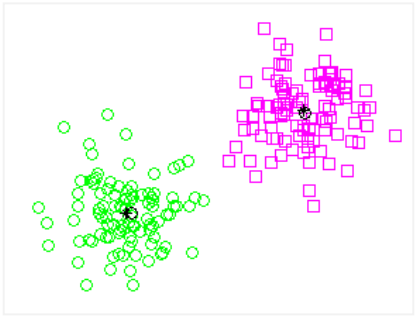
# Clustering applications 2/3

- Application: Document clustering
- Find groups of documents (topics) that are similar to each other based on the important terms appearing in them.
- Approach:
  - Identify important terms in each document.
  - Form a similarity measure between documents.
  - Cluster based on the similarity measure.
- Gain:
  - Help the end user to navigate in the collection of documents (based on the extracted clusters).
  - Utilize the clusters to relate a new document or search term to clustered documents.
- Check for example, Google News.

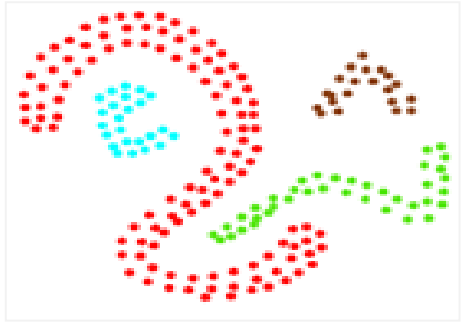
# Clustering applications 3/3



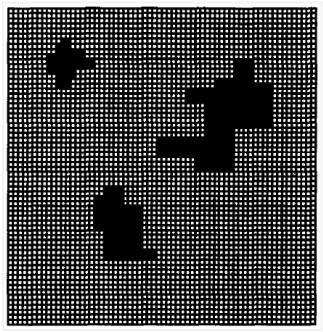
# A huge variety of clustering algorithms



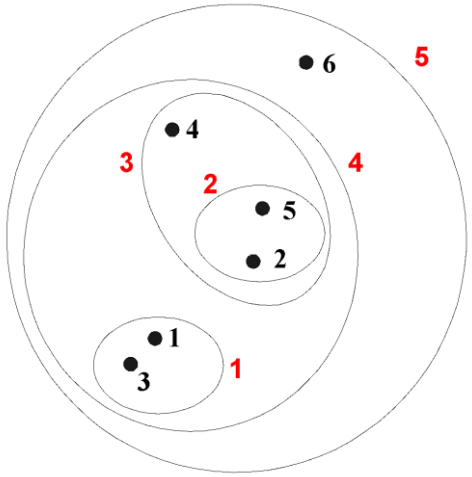
Partitioning methods
(*k*-Means)



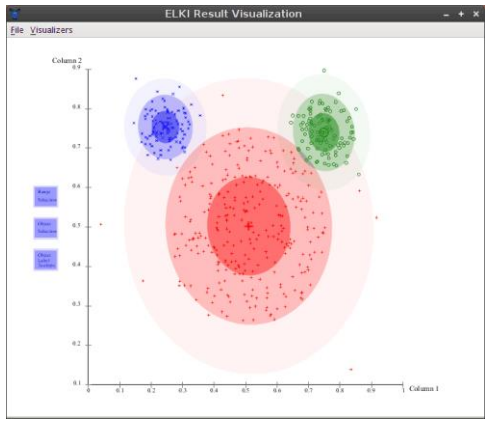
Model-based methods
(*DBSCAN*)



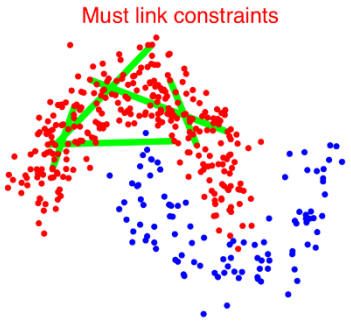
Grid-based methods
(*CLIQUE*)




Hierarchical methods

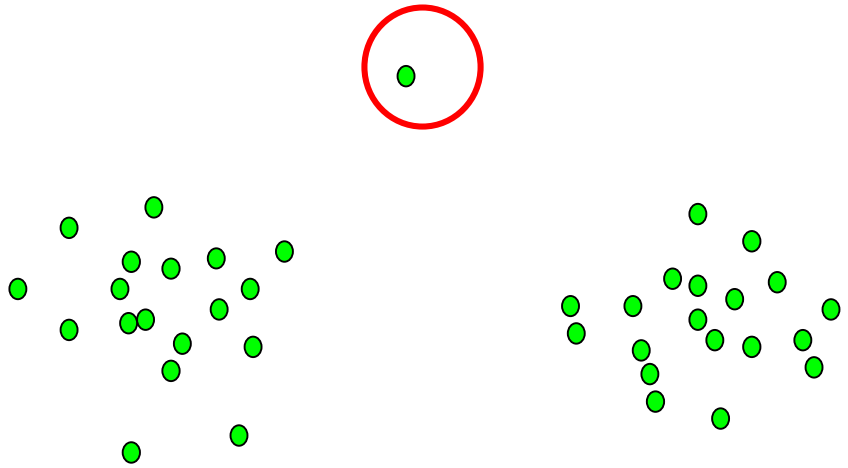



Model-based methods
(EM)




Constraint-based methods

# Learning tasks: Unsupervised/Supervised: Outlier detection



- Outlier detection is defined as identification of non-typical data
- Outliers might indicate
  - possible abuse of credit cards, mobile phones
  - data errors
  - device failures
- For some applications, outliers are more important than “normal patterns”
- Outliers might be global or local

# Application

- Analysis of the SAT.1-Ran-Soccer-Database (Season 1998/99)
  - 375 players
  - Primary attributes: Name, #games, #goals, playing position (goalkeeper, defense, midfield, offense),
  - Derived attribute: Goals per game
  - Outlier analysis (playing position, #games, #goals)
- Result: Top 5 outliers

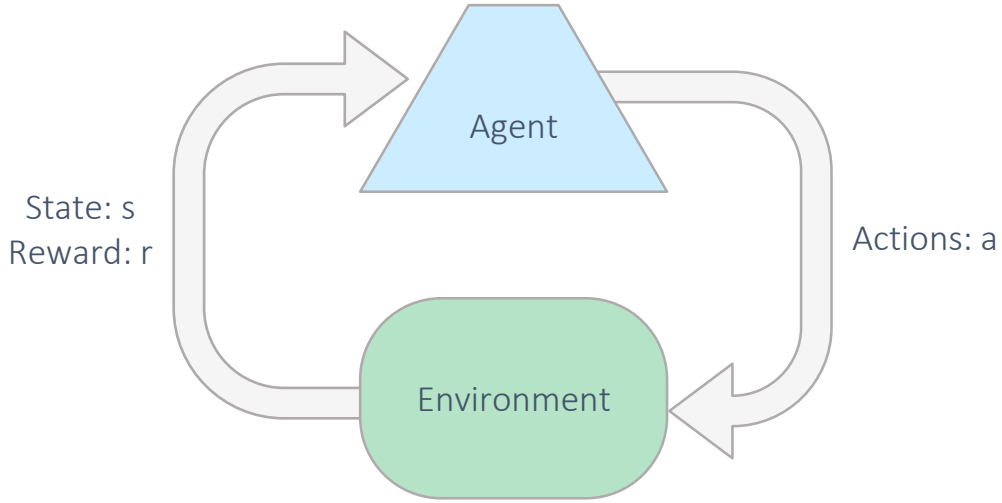
| Rank | Name | # games | #goals | position | Explanation |
|---|---|---|---|---|---|
| 1 | Michael Preetz | 34 | 23 | Offense | Top scorer overall |
| 2 | Michael Schjönberg | 15 | 6 | Defense | Top scoring defense player |
| 3 | Hans-Jörg Butt | 34 | 7 | Goalkeeper | Goalkeeper with the most goals |
| 4 | Ulf Kirsten | 31 | 19 | Offense | 2nd scorer overall |
| 5 | Giovanne Elber | 21 | 13 | Offense | High #goals/per game |

Note: “Outliers” is not necessarily a negative term.

# Learning tasks: Reinforcement learning

- RL is a type of ML technique that enables an agent to learn in an interactive environment by trial and error using feedback from its own actions and experiences.

- Basic idea:
  - Agent receives feedback in the form of rewards
  - Agent's utility is defined by the reward function
  - Must (learn to) act so as to maximize expected utility
  - All learning is based on observed samples of outcomes!

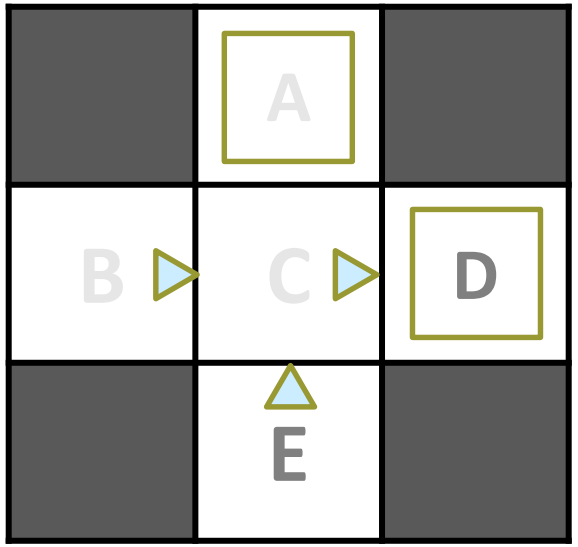



- Interaction: Modeled as a Markov Decision Process (MDP)

# Reinforcement learning: example

- Which action the agent should choose at each state?

Input Policy $\pi$




Observed Episodes (Training)

Episode 1

B, east, C, -1
C, east, D, -1
D, exit, x, +10

Episode 2

B, east, C, -1
C, east, D, -1
D, exit, x, +10

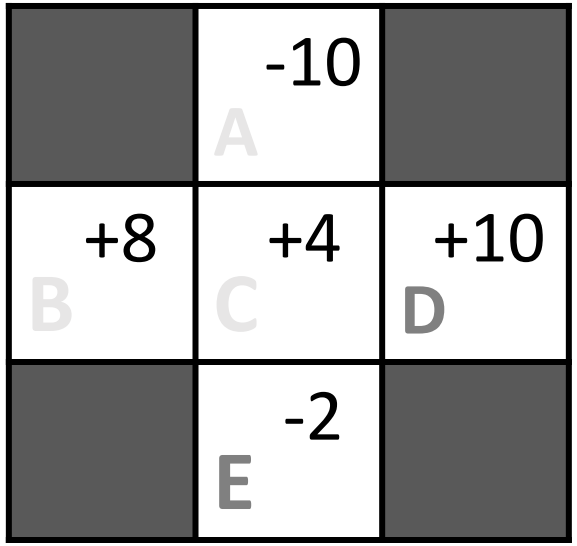
Episode 3

E, north, C, -1
C, east, D, -1
D, exit, x, +10

Episode 4

E, north, C, -1
C, east, A, -1
A, exit, x, -10

Output Values



# Outline

- Why to study Machine Learning/Data Science?
- Why we need Machine Learning?
- What is Machine Learning/ Data Science?
- Main (machine) learning tasks
- Course content & logistics
- Things you should know from this lecture & reading material

# What you will learn in this course?

- Introduction
- Part 1: Basic ML tasks
  - Supervised learning
  - Unsupervised learning
  - Reinforcement learning
  - Outlier detection
- Part 2: ML for particular/modern data challenges
  - High-dimensional learning
  - Learning over non-stationary data
  - Label/Data scarsity

# Part 1: Basic ML tasks

(Already introduced)

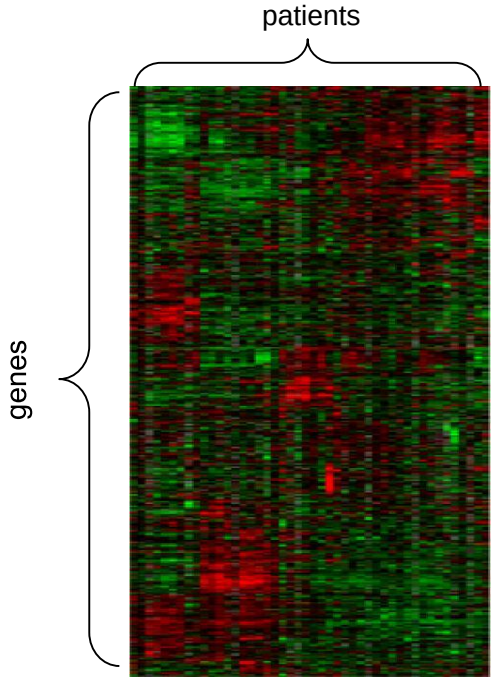
- Supervised learning
  - Classification
  - Regression
- Unsupervised learning
  - Clustering
- Reinforcement learning
- Outlier detection

- For each task, we will cover
  - Key methods and algorithms
  - Evaluation

# Part 2: Mind the data characteristics

- In the first part, several assumptions are made about the data
  - Data stationarity (the data distribution does not change)
  - i.i.d. distribution (all instances are independent and identically distributed)
  - Availability of labels (for supervised learning)
  - Low-dimensionality
  - …
- In the second part, we will waive some of these assumptions taking into account (real/modern) data challenges), namely
  - High-dimensionality
  - Non-stationarity
  - Label/Data scarcity
- We will discuss how ML methods (part 1) are affected by these challenges and new methods and algorithms for the particular data challenges

# High-dimensionality




- Many application generate high dimensional data
  - Text
  - Recommendation systems (e.g., Amazon products, IMDB movies, …)
  - Micro-array data
  - …
- Challenges when working with high-dimensional data
  - Distance measures loose their disriminative power in high dimensions (Curse of Dimensionality)
  - Patterns might occur in different subspaces and projections (each pattern might be only observable in certain subspaces)
  - …
- In this lecture we will focus on
  - Feature selection
  - Dimensionality reduction
  - Learning in subspaces (high dimensional clustering)

# Non-stationarity

- Most interesting applications nowadays come from *dynamic environments* where data are generated continuously over time (*stream data*)
- An example: Real-time network health monitoring
  - What are the profiles of normal connections?
  - What are the characteristics of attacks?

The characteristics of the (normal, attack) connections might change with time

*Source: http://www.networkworld.com/article/2366962/microsoft-subnet/spellbound-by-maps-tracking-hack-attacks-and-cyber-threats-in-real-time.htmls*

- Challenges for streaming data
  - Only 1 look at the data: complete history of data is often not available
  - Models must be updated: Changes in the data incur changes in the (machine learning) models
  - …
- In this lecture we will focus on
  - Stream classification
  - Stream clustering

# Label scarcity

- Big data small labels: Despite the big volume, big data do not come with label information
- Unlabelled data: Abundant and free
  - E.g., image classification: easy to get unlabeled images
  - E.g., website classification: easy to get unlabeled webpages
- Labelled data: Expensive and scarce

- Goal: leverage unlabeled data together with the labeled ones to produce better models
- In this lecture we will focus on
  - Self-learning, co-learning, semi-supervised learning
  - Data augmentation
  - Synthetic data generation

# Course requirements

We assume knowledge of basics from

- Algorithms
- Statistics
- Probabilities
- Linear algebra
- Optimization
- …

but most of the things you need to know will be covered.

For the assignments/projects we assume knowledge of

- Python

# Tentative schedule

| Week | # | Day | Date | Task | Topic |
|---|---|---|---|---|---|
| 0 | 0 | Wedneday | 20/10/2021 | | |
| | 0 | Thursday | 21/10/2021 | | |
| 1 | 1 | Wedneday | 27/10/2021 | Data | Introduction |
| | 2 | Thursday | 28/10/2021 | | Getting to know your data (Exploratory analysis and features) |
| 2 | 3 | Wedneday | 03/11/2021 | Supervised learning | Intro & Decision Trees |
| | 4 | Thursday | 04/11/2021 | | KNNs & Evaluation |
| 3 | 5 | Wedneday | 10/11/2021 | | BN and NBs |
| | 6 | Thursday | 11/11/2021 | | SVMs |
| 4 | 7 | Wedneday | 17/11/2021 | | Linear regression & logistic regression |
| | 8 | Thursday | 18/11/2021 | | NNs |
| 5 | 9 | Wedneday | 24/11/2021 | Unsupervised learning | Partitioning-based methods |
| | 10 | Thursday | 25/11/2021 | | Hierarchical-based methods |
| 6 | 11 | Wedneday | 01/12/2021 | | Density-based methods |
| | 12 | Thursday | 02/12/2021 | | Evaluation |
| 7 | 13 | Wedneday | 08/12/2021 | RL | MDPs 1/2 |
| | 14 | Thursday | 09/12/2021 | | MDPS 2/2 |
| 8 | 15 | Wedneday | 15/12/2021 | | RL 1/2 |
| | 16 | Thursday | 16/12/2021 | | RL 2/2 |
| 9 | 17 | Wedneday | 05/01/2022 | Potpouri | Outlier detection |
| | 18 | Thursday | 06/01/2022 | | Wrapping up basic part, Q&As & advanced part intro |
| 10 | 19 | Wedneday | 12/01/2022 | High dimensionality | Dimensionality reduction / representation learning 1/2 |
| | 20 | Thursday | 13/01/2022 | | Dimensionality reduction / representation learning 1/2 |
| 11 | 21 | Wedneday | 19/01/2022 | | Learning in subspaces 1/2 |
| | 22 | Thursday | 20/01/2022 | | Learning in subspaces 2/2 |
| 12 | 23 | Wedneday | 26/01/2022 | Volatility | Stream classification 1/2 |
| | 24 | Thursday | 27/01/2022 | | Stream classification 2/2 |
| 13 | 25 | Wedneday | 02/02/2022 | | Stream clustering 1/2 |
| | 26 | Thursday | 03/02/2022 | | Stream clustering 1/2 |
| 14 | 27 | Wedneday | 09/02/2022 | Label Scarsity | Synthetic data generation 1/2 |
| | 28 | Thursday | 10/02/2022 | | Synthetic data generation 2/2 |
| 15 | 29 | Wedneday | 16/02/2022 | Wrapping up, Responsibility aspects | |
| EXAM | 30 | Thursday | 17/02/2022 | | |

# Course logistics: general

For in-person attendance, follow the official rules: https://www.mi.fu-berlin.de/w/Mi/LehreMitPraesenz

- ECTS points: 10
- Lectures
  - Wednesdays, 16:00-18:00, hybrid (Room: T9 (Takustr. 9)/Gr.Hörsaal & Webex link).
  - Thursdays, 12:00-14:00, hybrid (Room: A3 (Arnimallee 3)/Hs 001 Hörsaal & Webex link).
- Tutorials (2 groups, same content)
  - Tuesdays, 12:00-14:00 in-person (Room: T9 (Takustr. 9)/SR 005) & 14:00-16:00 (Webex Link)
- Exam
  - Planned for 17/2/2022
  - Form: online most probably

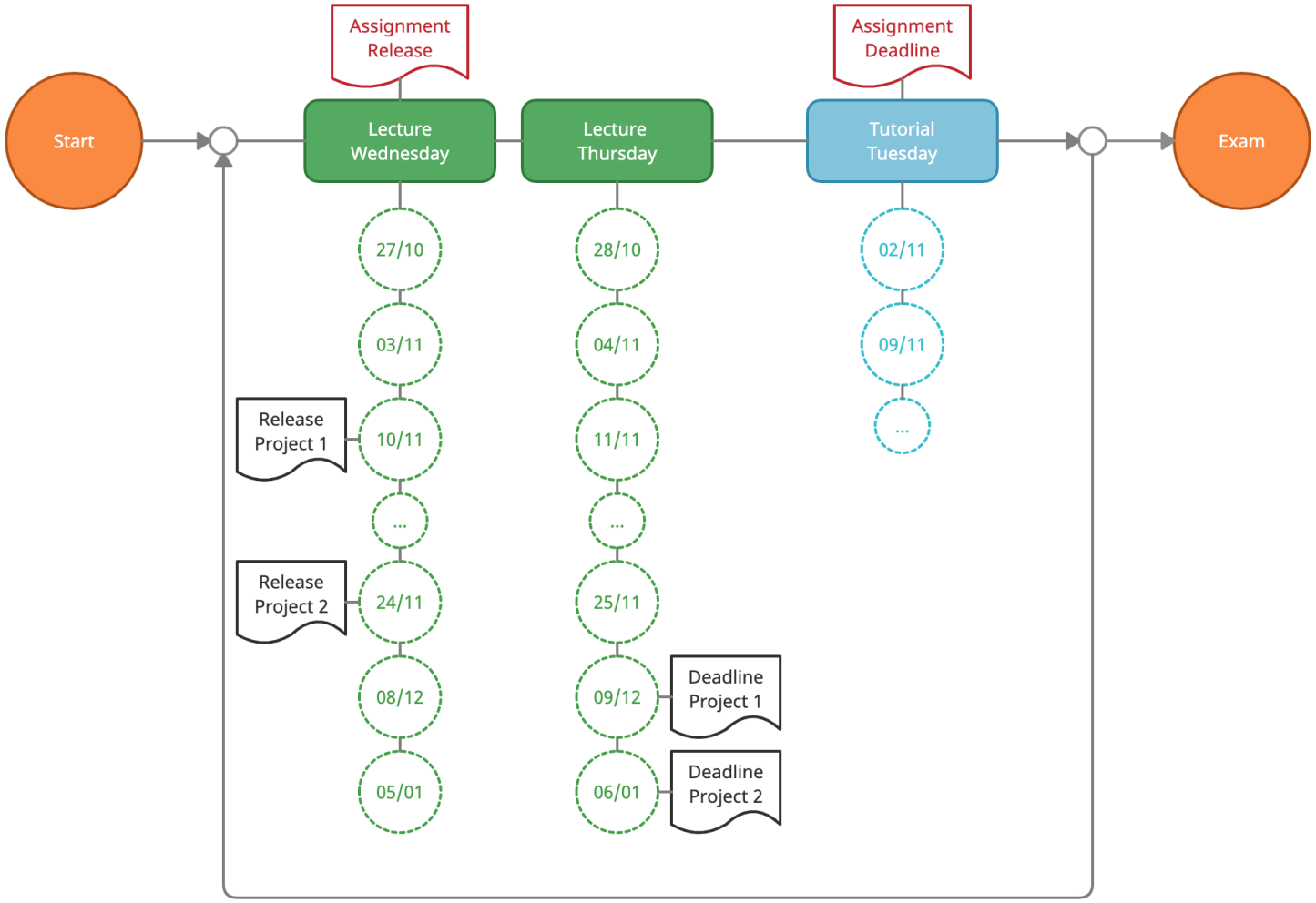
The current plan is a hybrid lecture, for which some students attend *in the classroom* and others *online* via livestreaming (Webex). Regarding the in-person teaching part, we remain flexible to the pandemic, student and teaching staff needs and we will make adjustments and changes accordingly.

- Up to date information on Whiteboard. Please check regularly/activate notifications!

# Course logistics: assignments, projects and active participation requirement

- 1 assignment sheet per week (announced after the lecture, the exact timeline will be discussed) → ~ n=14 sheets
- 3 projects on supervised learning (P1); unsupervised learning (P2); TBA (P3)
- Active participation requirement: You should pass
  - #n-4
  - ≥ 2 projects

# Timeline for the assignments and projects



# Working together

- We encourage cooperation and exchange between students
- We recommend using Mattermost as an exchange space for posting questions and answering to questions of other students
  - Having all the Q&A in one place is beneficial for everyone.
  - We will monitor Mattermost for open questions and will address them in lectures/tutorials.
  - Questions received per email will not be answered.

- Groups of 2-3 students
  - We dont interfere with group formation, how you split the work within the group, etc.
  - Every group member receives the same grade (Please name all group members for each submission in the Whiteboard text field)

# Course logistics: team

- Lectures: Prof. Dr. Eirini Ntoutsi

*Eirini*

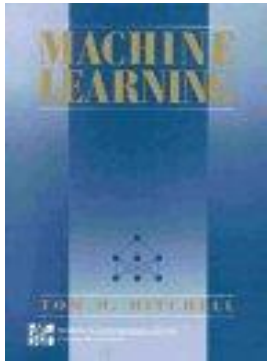
- Tutorials:
  - Main TA: M.Sc. Manuel Heurich

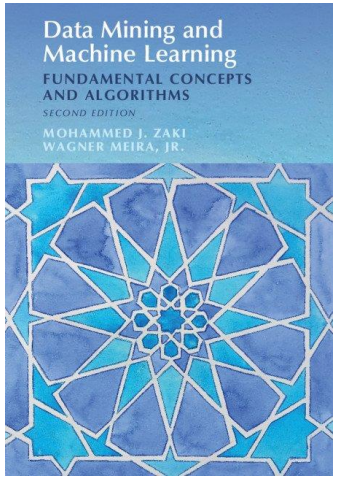
*Manuel*

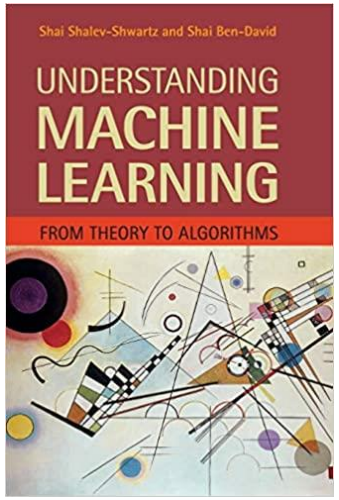
- Contributing in slides/assignement material
  - M.Sc. Arjun Roy
  - M.Sc. Emanouil Panagiotou
  - M.Sc. Philip Naumann
  - M.Sc. Yi Cai
  - …

# Recommended readings

- Many great books! Below, no preference order is implied
- Mitchell T. M., Machine Learning, McGraw-Hill, 1997
  - http://www.cs.cmu.edu/~tom/mlbook.html
- Meira and Zaki, Data Mining and Machine Learning: Fundamental Concepts and Algorithms, Cambridge University Press, 2020
  - https://dataminingbook.info/
- Shai Ben-David and Shai Shalev-Shwartz, Understanding Machine Learning: From Theory to Algorithms, Cambridge University Press, 2014
  - https://www.cambridge.org/core/books/understanding-machine-learning/3059695661405D25673058E43C8BE2A6




- Slides should be sufficient though
- Further reading material might be provided for each lecture

# Overview and Reading

- Overview
  - What is ML?
  - Why we need ML?
  - What are main ML tasks?
  - (Some) Modern data challenges

- Reading
  - Introductory chapter from your favorite book
  - Check resources linked in the slides

# Next assignment and tutorial

- Next (first) assignment to be announced directly after the lecture today
  - Assignment announcement via Whiteboard
  - You need to submit your solution to the same assignment entry (only PDFs)
    - Your coding solutions should be exported as PDF before uploading them
  - Submission deadline: Tuesday (02/11) at noon (before the tutorial starts)

- Next (first) tutorial on Tuesday next week
  - Going through the solutions of the tasks that are due until Tuesday
  - Introducing the next tasks that will be due the week after
  - Answering upcoming questions

# Thank you

Questions/Feedback/Wishes?

# Acknowledgements


- The slides are based on
  - KDD I lecture at LMU Munich (Johannes Aßfalg, Christian Böhm, Karsten Borgwardt, Martin Ester, Eshref Januzaj, Karin Kailing, Peer Kröger, Eirini Ntoutsi, Jörg Sander, Matthias Schubert, Arthur Zimek, Andreas Züfle)
  - Introduction to Data Mining book slides at http://www-users.cs.umn.edu/~kumar/dmbook/
  - Pedro Domingos Machine Lecture course slides at the University of Washington
  - Machine Learning book by T. Mitchel slides at http://www.cs.cmu.edu/~tom/mlbook-chapter-slides.html
  - Thank you to all TAs contributing to their improvement, namely Vasileios Iosifidis, Damianos Melidis, Tai Le Quy, Han Tran